# Dicionário-Aberto: Construção semiautomática de uma funcionalidade codificadora

## *1. Introdução*

O nosso projeto visa transformar o Dicionário Aberto (DA) num avançado dicionário codificador ou de produção. O que se pretende é que, utilizando as funcionalidades de pesquisa reversa do Dicionário Aberto, o utilizador possa procurar unidades lexicais relacionadas (sinónimos, quase-sinónimos, hiperónimos, hipónimos, merónimos, holónimos, co-ocorrentes, etc.) a partir de uma ou de várias palavras.

Para que isto seja possível, será necessário que o sistema de pesquisa disponibilizado aos utilizadores não use apenas os termos introduzidos pelo utilizador e as respectivas ocorrências nas definições, mas também uma estrutura semântica que providencie relações léxico-conceptuais com os termos introduzidos. Estas relações serão cruzadas e serão calculadas medidas de proximidade, sendo deste modo possível apresentar um conjunto de resultados ordenados por relevância.

Na secção 2 é apresentado o projeto, a sua origem e um pouco da sua história. De seguida, na secção 3, são apresentadas as funcionalidades básicas de pesquisa presentes em praticamente todos os dicionários disponíveis em linha. Por sua vez, a secção 4 apresenta as funcionalidades avançadas de pesquisa, como sejam a pesquisa reversa ou a pesquisa ontológica. Finalmente, na secção 5 são tiradas algumas ilações referentes ao trabalho desenvolvido.

## *2. O Dicionário Aberto*

O DA (disponível na rede, para consulta e para extração automática de informação, em http://www.dicionario-aberto.net, mas também para uso local, de modo aberto e gratuito) iniciou em junho de 2005, como a transcrição da edição de 1913 dos dois volumes do *Novo Diccionário da Língua Portuguesa,* de Cândido de Figueiredo. Para a transcrição do dicionário usou-se uma aplicação *web* criada para ajudar na transcrição de livros para integrarem o acervo do *Projecto Gutenberg* (http://www.gutenberg.org/). Esta aplicação está disponível num sítio da Internet, chamado PGDP (*Project Gutenberg, Distributed Proofreaders,* http://www.pgdp.net) (Newby & Franks, 2003), onde é apresentado o texto de uma página, resultante de um OCR (*Optical Character Recoginition*) que o utilizador deve ler e corrigir de acordo com a imagem disponibilizada.

O processo de transcrição, efetuado inteiramente por voluntários, terminou em 2010. Durante estes cinco anos o sítio *web* do dicionário tornou forma, e foi incorporando, diariamente, um conjunto de novas palavras que terminavam a passagem pelas várias rondas de revisão.

Desde então o dicionário tem sido sujeito a um conjunto de alterações, como a transformação num formato standard e sintaticamente rico (TEI - Text Encoding Initiative - http://www.tei-c.org/), a validação da notação e, mais recentemente, a modernização da ortografia (Simões & Farinha, 2011).

Embora esta última etapa não esteja concluída, por necessitar de validação manual do resultado da ferramenta automática de modernização, o dicionário está disponível livremente não só para a pesquisa em-linha, que será seguidamente apresentada, mas também em diferentes formatos para que possa ser usado em diferentes contextos. Destes vários formatos salientamos o formato StarDict (http://www.stardict.org), usado por exemplo na ferramenta de tradução assistida por computador OmegaT (http://www.omegat.org), o formato SQL (Structured Query Language), para poder ser incorporado numa base de dados local e processada computacionalmente, ou ainda a disponibilização através de um serviço RESTful (Fielding 2000) que permite, por exemplo, o desenvolvimento de aplicações móveis do DA que consultem dinamicamente a versão mais recente disponível.

## *3. A pesquisa simples*

A aplicação web do DA tem evoluído, incorporando diversos tipos de pesquisa: pela entrada do artigo, partes do lema (início, meio ou fim) ou ainda a pesquisa reversa, entre outras funcionalidades. Mais recentemente incorporou-se um sistema semiautomático para a extração de sinónimos/antónimos, hiperónimos/hipónimos e merónimos/holónimos que serve para explorar as relações léxico-conceptuais entre os termos introduzidos na pesquisa (Simões, Iriarte & Almeida, 2012).

Na pesquisa simples, o utilizador já vai encontrar palavras relacionadas com a entrada procurada (como sejam sinónimos, hiperónimos ou merónimos). A discussão sobre como estas palavras são obtidas e os seus potenciais problemas é feita na secção 4.

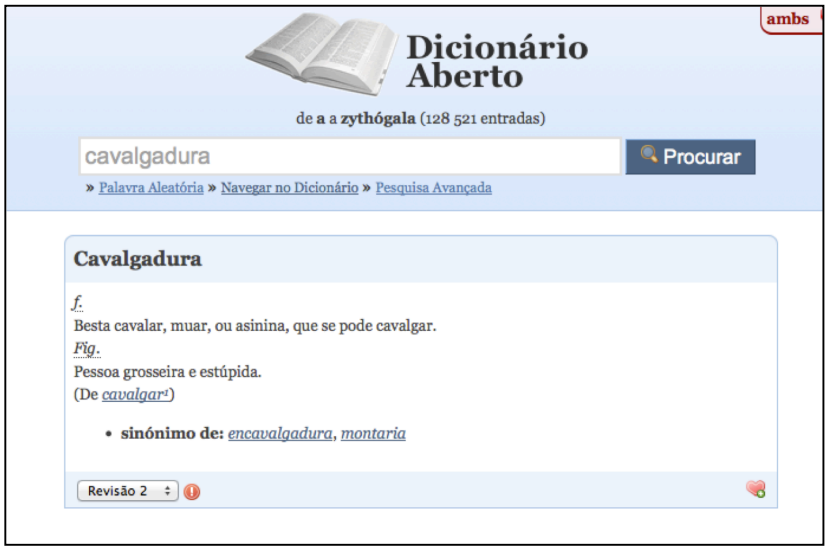


**Figura 1**

Resultado da pesquisa simples da palavra "cavalgadura".

A pesquisa simples também permite obter um conjunto de palavras ortograficamente semelhantes (Levenshtein, 1966), como mostra a figura 2, úteis para quando se desconhece a grafia exata de um lema.

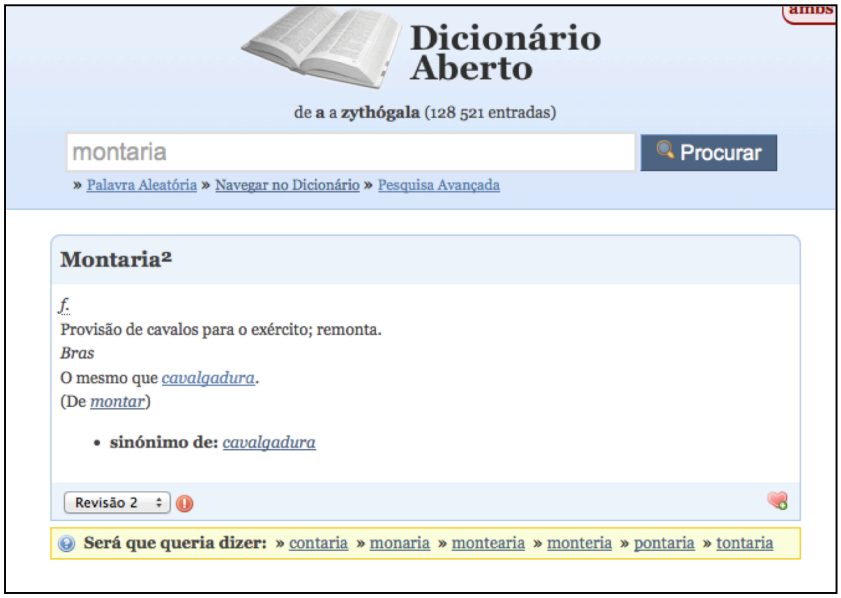


**Figura 2**

Procura da palavra "montaria" e apresentação de palavras ortograficamente semelhantes.

Finalmente, existe uma outra funcionalidade que permite ao utilizador navegar pelas entradas do dicionário de forma sequencial, vendo as palavras ortograficamente próximas ordenadas alfabeticamente. Esta abordagem tenta assemelhar-se ao processo de folhear um dicionário. A figura 3 apresenta esta funcionalidade. Repare-se na possibilidade de obter a palavra seguinte ou anterior (presentadas acima do verbete, com setas para a esquerda e para a direita ) e na lista de palavras anteriores e posteriores à entrada atual, apresentada na coluna à esquerda.

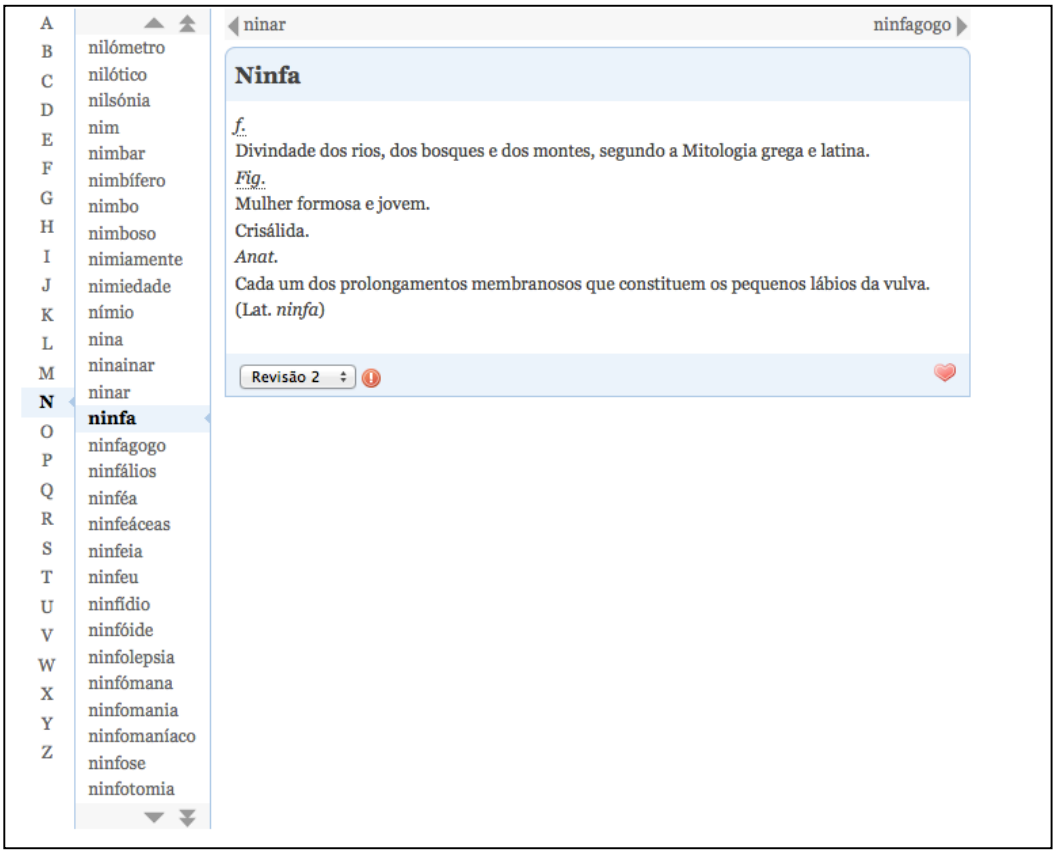


**Figura 3**

Verbete da palavra "ninfa" com a interface de navegação.

Embora estas funcionalidades sejam úteis para grande parte das situações em que um utilizador consulta um dicionário convencional, é possível adicionar outras que tornem o dicionário muito mais útil, funcional e poderoso, como veremos na secção seguinte.

## *4. A pesquisa avançada*

Muito mais interessantes são as funcionalidades avançadas de pesquisa, que transformam o DA num verdadeiro dicionário codificador. Utilizando as funcionalidades de pesquisa reversa ou pesquisa ontológica, por exemplo, o utilizador poderá procurar unidades lexicais relacionadas (sinónimos, quase-sinónimos, hiperónimos, hipónimos, merónimos, holónimos, co-ocorrentes, etc.) a partir de uma ou de várias palavras. Estas funcionalidades avançadas, disponíveis para utilizadores registados, estão a transformar o DA numa ferramenta muito útil para linguistas e investigadores em Processamento da Linguagem Natural.

### *4.1. A pesquisa por prefixo, infixo e sufixo*

Após ter selecionado a "pesquisa avançada", é possível pesquisar fragmentos de lema (início, meio ou fim da palavra), selecionando na interface a palavra "Prefixo", "Infixo" ou "Sufixo". Estes termos, que esperamos poder vir a substituir, não são os mais felizes, uma vez que os resultados não correspondem a estas categorias morfológicas, mas são suficientemente transparentes e, principalmente, curtos para poderem ser utilizados confortavelmente na interface gráfica e entendíveis pelo público em geral.

Esta funcionalidade pode ser uma ferramenta importante para a morfologia (Millán, 1999). Imagine-se, por exemplo, a sua utilidade para o estudo da produtividade dos afixos, como a dos diminutivos em *-inho, -ito - ino*, etc.; a produtividade de determinados sufixos na terminologia científica: *-ato, -eto - ito*; da produtividade e verdadeira sinonímia ("ação ou resultado/efeito de") de sufixos como *–dade/ –ção/ –são, –ança/ –ância*; etc.

O DA pode vir a ser um recurso importante para a investigação linguística e para a elaboração gramáticas e de outros dicionários, uma vez que nos permite descarregar os resultados destas pesquisas. Por exemplo, a figura 4 apresenta um estudo da produção de advérbios em *-mente* a partir de adjetivos em *-vel.*

| | |
|---|---|
| ... | ... |
| Agitável | - |
| Aglutinável | - |
| Agradável | Agradavelmente |
| Agradecível | - |
| Agricultável | - |
| Ajuntável | - |
| Alcançável | - |
| Alcoolizável | - |
| Alheável | - |
| Aliável | - |
| Alienável | - |
| Alliável | - |
| Alterável | - |
| Amável | Amavelmente |
| Amigável | Amigavelmente |
| Amissivel | - |
| Amoedável | - |
| Amoldável | - |
| Amolgável | - |
| Amorável | Amoravelmente |
| Amortizável | - |
| Amotinável | - |
| Amovível | - |
| Amparável | - |
| ... | ... |

**Figura 4**

Análise do léxico tendo por base listas de palavras descarregadas a partir do Dicionário-Aberto.

A pesquisa por sufixos pode também ser usada a modo de dicionário de rimas gráficas. Também é fácil incluir a capacidade de pesquisa inversa: transformando o DA num dicionário inverso, em que a ordenação alfabética dos lemas é feita da direita para a esquerda.

*4.2. A pesquisa reversa*

A pesquisa reversa transforma os dicionários eletrónicos numa espécie de base de dados conceptual, que funciona a modo de dicionário onomasiológico ou de dicionário de produção ou codificador.

Utilizando as funcionalidades de pesquisa reversa do DA, o utilizador pode procurar unidades lexicais relacionadas (sinónimos, quase-sinónimos, hiperónimos, hipónimos, merónimos, holónimos, co-ocorrentes, etc.) a partir de um conjunto de palavras. Um exemplo das potencialidades como dicionário codificador ou dicionário de produção: introduzindo nas janelas de pesquisa as palavras "endurecer" e "metal" obtemos como resultado a entrada "temperar" (ver figura 5).

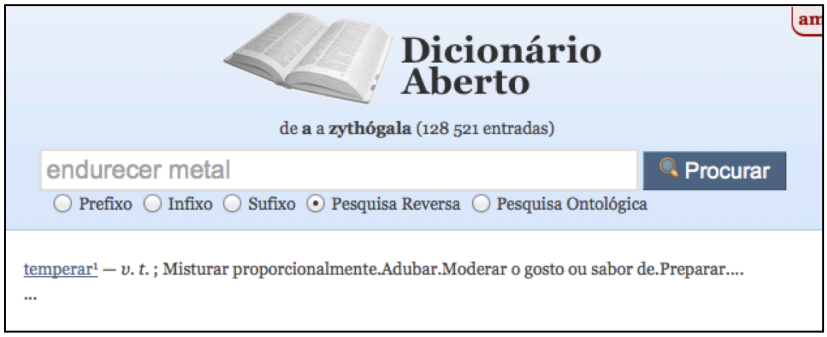


**Figura 5**

Pesquisa reversa das palavras "endurecer" e "metal", obtendo o resultado "temperar".

Esta pesquisa é efetuada sobre as definições e não sobre os lemas, o que acaba por ser uma limitação, já que só serão encontradas entradas que usem exatamente as palavras usadas na pesquisa. A esse respeito existe um conjunto de melhorias a serem implementadas a este tipo de pesquisa, nomeadamente:

– o uso de um analisador morfológico para que não sejam indexadas as formas das palavras ocorrentes das definições, mas os seus lemas, e para que as palavras procuradas sejam também lematizadas antes de serem usadas na pesquisa. Embora esta funcionalidade seja relativamente fácil de implementar levará a que existam alguns problemas na cobertura do analisador morfológico;

– as palavras pesquisadas podem ocorrer em qualquer sítio da definição. Seria bastante interessante que fosse possível procurar por uma sequência de palavras exata, tal como acontece com os motores de pesquisa na Internet.

### *4.3. A pesquisa ontológica*

Muito mais interessante e prometedor é o que chamamos de "pesquisa ontológica". Esta pesquisa é baseada numa ontologia construída dinamicamente (e portanto, de forma completamente automática). De seguida resume-se a abordagem utilizada na extração automática da ontologia a partir das definições (Simões, Álvaro a Almeida, 2012), e posteriormente é explicado o algoritmo de pesquisa baseado nesta ontologia.

Para a extração automática de relações léxico-conceptuais usa-se um conjunto de regras ou padrões (Hearst 1992). Estes padrões são sequências de palavras que se pretendem encontrar nas definições e que indicam a grande probabilidade da palavra que segue o padrão estar relacionada com o lema da respetiva definição. Isto só é possível graças à regularidade da estrutura das definições.

Seguem-se alguns exemplos de regras (uma por relação léxico-conceptual) que foram usadas na extração de relações:

– Sinonímia (SYN): *o mesmo (ou melhor) que ...* ;

– Antonímia (ANT): *que não é ...* ;

– Hiponímia (HIPO): *espécie de ...* ;

– Hiperonímia (HIPER): *que tem por tipo...* ;

– Meronímia (MERO): *cada uma das partes que formam ...* ;

– Holonímia (HOLO): *composto por...* .

Para esta funcionalidade, também se usam relações calculadas (por exemplo, utilizando a transitividade da relação de hiperonímia). Somos conscientes de que algumas regras podem ser problemáticas (por exemplo, no caso dos sinónimos e quase-sinónimos). Em todo o caso, parece-nos preferível apresentar um conjunto de possíveis falsos sinónimos do que garantirmos a correção e diminuir drasticamente o número de relações da ontologia resultante. Note-se que uma ontologia torna-se mais útil quando há grande diversidade nos relacionamentos existentes.

Assim, definiram-se regras matemáticas para a "completação" da ontologia, inferindo novos relacionamentos a partir dos relacionamentos iniciais:

– a SYN b ⇒ b SYN a – propriedade de simetria entre sinónimos: se *a* é sinónimo de *b,* então *b* também é sinónimo de *a*. Esta relação é bastante produtiva já que em muitas situações de verdadeiros sinónimos o dicionarista definiu a relação de sinonímia apenas num dos verbetes das palavras envolvidas;

– a HIPO b ∧ c HIPO b ⇒ a COHIPO c – a relação de co-hiponímia não pode ser obtida diretamente usando padrões de Hearst, mas pode ser calculada a partir das relações de hiponímia. Assim, se duas palavras *a* e *c* são hipónimos da mesma palavra *b,* então *a* e *c* são co-hipónimos. Esta relação é extremamente útil já que tipicamente relaciona objetos de um mesmo tipo ou classe;

– a HIPO b ∧ b HIPO c ⇒ a HIPO c – a transitividade das relações hierárquicas, como é o caso da hiperonímia ou hiponímia, permite que se possa procurar termos genéricos com base num termo bastante específico ou termos específicos usando termos bastante genéricos. Como exemplo prático, considere-se a pesquisa pelo termo "animal". É pouco provável que se encontrem entradas referentes a animais diretamente relacionados com este termo, mas por transitividade poderemos encontrar outras classes como "mamífero" ou "peixe" como hipónimos de "animal", e como hipónimos destes, encontrar entradas correspondentes a animais.

Na pesquisa ontológica, os resultados da pesquisa são as entradas que se relacionam com essas palavras (ordenadas por quantidade de relacionamentos). Repare-se que, ao contrário do que acontece com a pesquisa reversa, alguns dos resultados são entradas que não contêm nenhum dos termos introduzidos pelo utilizador na janela de pesquisa. Assim, no exemplo recolhido na figura 5, o terceiro resultado, entre outros, corresponde a uma entrada que não contém os termos introduzidos pelo utilizador ("veado" e "animal").

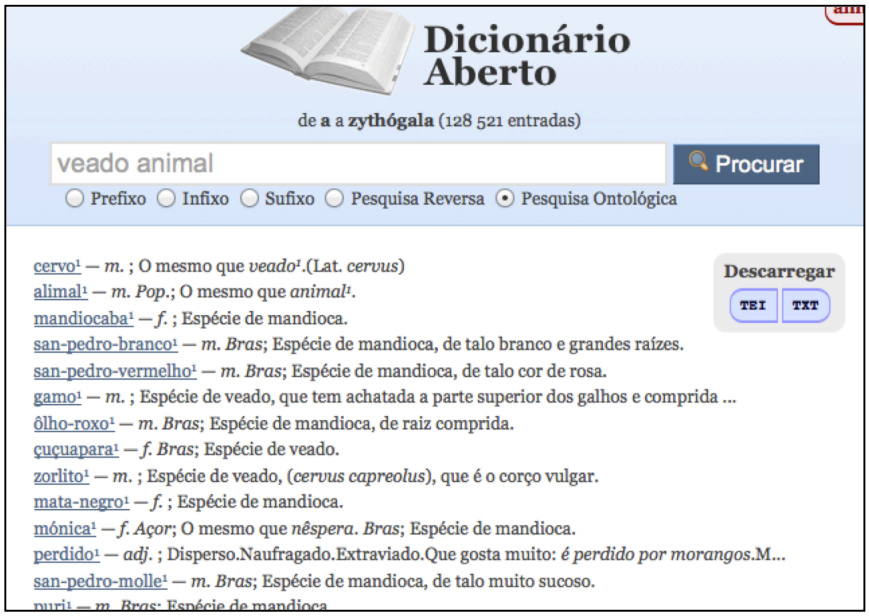


**Figura 6**

Pesquisa ontológica a partir dos termos "veado" e "animal".

Tal como no caso da pesquisa reversa, a pesquisa ontológica pode beneficiar de algumas novas funcionalidades. Se por um lado a ontologia extraída poderá lucrar em ser expandida usando, por exemplo, uma ontologia externa como o Onto.PT (Oliveira & Gomes, 2012), por outro lado os termos de pesquisa deveriam ser lematizados antes de serem usados no algoritmo de procura. Além disso, a possibilidade de incorporar funcionalidades de pesquisa reversa juntamente com as funcionalidades de pesquisa ontológica seria muito proveitosa.

## *4. Conclusões*

Estamos convictos de que o DA é uma excelente ferramenta que pode ser usada como um dicionário tradicional, mas também como um recurso para tarefas de processamento de linguagem natural e como ferramenta para ajudar na verificação de hipóteses colocadas na investigação linguística e na elaboração gramáticas e de outros dicionários.

O facto de ser um recurso aberto poderá também significar que as suas funcionalidades poderão vir a aumentar em quantidade e em qualidade. Mas, por outro lado, o facto de não ser um projeto financiado limita a possibilidade de contratação de mão de obra especializada para a revisão exaustiva, por exemplo, da modernização da língua. Ou seja, a evolução da qualidade das entradas do DA está limitada à disponibilidade de voluntários.

Por outro lado, e embora a evolução do conteúdo linguístico do dicionário seja lenta, a implementação de algoritmos e a execução de experiências sobre o DA tem sido bastante proveitosa, demonstrando que é possível criar funcionalidades úteis a partir de dicionários convencionais.

Universidade do Minho - Portugal

Alberto SIMÕES

Álvaro IRIARTE

José João ALMEIDA

## *Referências bibliográficas*

Almeida, José João / Santos, André / Simões, Alberto, 2010. «Bigorna – a toolkit for orthography migration challenges», in : Calzolari, N., et al., (ed.), *Seventh International Conference on Language Resources and Evaluation (LREC 2010),* Valletta, Malta, 227–232.

Fielding, Roy T. 2000. «Representational State Transfer (REST)», *Architectural Styles and the Design of Network-based Software Architectures*, PhD Dissertation, Irvine, University of California.

Hearst, Marti, 1992. «Automatic acquisition of hyponyms from large text corpora», *Proceedings of the Fourteenth International Conference on Computational Linguistics,* Nantes, France, Volume 2, 539–545.

Levenshtein, Vladimir Iosifovich, 1966. «Binary codes capable of correcting deletions, insertions, and reversals», *Soviet Physics Doklady* 10, 707–10.

Millán, José Antonio, 1999. «Zigzag, gong, ping-pong, iceberg. donde se descubre que hay diccionarios inversos, y su utilidad manifiesta para el progreso de la humanidad» <http://jamillan.com/inverso.htm> (último acesso em 1 de julho de 2013).

Millán, José Antonio, 2011. «Dirae: consulte el DRAE como ya no podía hacerlo» <http://jamillan.com/librosybitios/2011/05/dirae-consulte-el-drae-como-ya-no-podia-hacerlo/> (último acesso em 1 de julho de 2013).

Newby, Gregory B. / Franks, Charles, 2003. «Distributed proofreading», *Proceedings of the Joint Conference on Digital Libraries,* 361–363.

Oliveira, Hugo Gonçalo / Gomes, Paulo, 2012. «Ontologising semantic relations into a relationless thesaurus», *Proceedings of 20th European Conference on Artificial Intelligence (ECAI 2012)*, Montpellier, France, August 2012. IOS Press, 915-916.

Simões, Alberto / Farinha, Rita, 2011. «Dicionário Aberto: Um novo recurso para PLN», *Vice-versa* 16, 159–171.

Simões, Alberto / Iriarte, Álvaro / Almeida, José João, 2012 «Dicionário-aberto – a source of resources for the portuguese language processing», in : Caseli, H. / Villavicencio, A. / Teixeira, A., / Perdigão, F. (ed.) *Computational Processing of the Portuguese Language, Lecture Notes for Artificial Intelligence,* Berlim, Springer, 7243, 121–127.